Journal of Applied Data Sciences
Vol. 7, No. 2, May 2026, pp.1265-1280
ISSN 2723-6471
1265

Predicting 2000-Meter Indoor Rowing Performance Using Accessible Machine Learning Models

Arihant Jaggi[1,*], Hiren Dandia[2]

[1]The Cushman School, Miami, Florida, USA 33137

[2]Mentor, Pangea Society, Gurgaon, Haryana, India 122002

(Received: December 1, 2025; Revised: January 25, 2026; Accepted: March 25, 2026; Available online: April 18, 2026)

Abstract

The 2000-meter ergometer test is widely used to measure athlete's strength, skill and efficiency in competitive rowing. Traditional tests like 500m or 1000m rowing can be too physical and elaborate for beginners or younger rowers. This study aimed to create a simpler and data driven approach to predict 2000m rowing times using basic information like age, gender and weights. Predictions were made using machine learning models including XGBoost that were applied to data from 1,341 rowers obtained from Concept 2 Database and Miami Beach Rowing Club. The model performed better for athletes over 18 years old with gender as the most important factor followed by weight and age. Finally after rigorous model training, the model showed insightful prediction accuracy with R2=0.75, MAE=0.35 min and RMSE=0.47 min. However, cross-validation of the model showed R2=-2.04, indicating overfitting due to limited variables and data. Despite this limitation, our model offers a practical application that can help rowers set realistic goals and assist coaches in personalized training. In conclusion, the model can still be improved to improve accuracy and validation but in the current study it represents a step forward in making performance insights more accessible to rowers.

Keywords: Rowing, 2000-Meter Ergometer, AI Prediction, Machine Learning, Xgboost

1. Introduction

The 2000-meter ergometer test is used as a standard for testing an athlete's physical fitness and strength in rowing [1]. It offers insights into rower's skills, speed and efficiency for success in rowing. Accurately predicting 2000-meter times goes beyond being just an academic exercise; it provides athletes with concrete, realistic performance targets that are crucial for motivation and structured training. Additionally, it equips coaches with invaluable data to optimize training regimens, make informed recruitment decisions and strategically plan for the 2000-meter race [2].

Historically, performance estimation in rowing has relied on strenuous physical tests such as repeated 500-meter [3] or 1000-meter intervals [4]. While these traditional methods are recognized for their accuracy and benefits in assessing an athlete's current condition, they also come with significant practical drawbacks. These tests can be time-consuming and demanding, requiring substantial recovery periods and maximal exertion. This challenging nature limits their widespread applicability, particularly for developing athletes, younger individuals or during the early stages of training when intense physical exertion might be counterproductive or even pose health risks. Consequently, there is a substantial gap between the ongoing need for performance forecasting and the practical feasibility of traditional assessment methods.

To address this gap, this study introduces a transformative approach that utilizes machine learning and artificial intelligence. The core innovation lies in its strategic focus on readily available and consistently measurable variables: age, gender and weight. This selection directly addresses the accessibility limitations of conventional methods, as these data points are typically easy to obtain for athletes without requiring specialized equipment, laboratory settings or maximal physical effort. This fundamental shift in data collection and assessment significantly lowers the entry barrier for performance analysis. By making advanced predictive analytics feasible for a broader population of athletes and

***Corresponding author: Arihant Jaggi (s.ajaggi@cushmanschool.org)**
DOI: https://doi.org/10.47738/jads.v7i2.999

 ISSN 2723-6471

coaches, the utility of performance prediction in rowing is effectively expanded. This move shifts performance assessment from being an exclusive domain of elite athletes with access to specialized facilities to one accessible to a wider range of participants, fostering broader engagement and development within the sport.

The timing of this research is particularly opportune, benefiting from two converging trends. Firstly, the increasing digitization of sports data, exemplified by extensive databases like Concept 2 [5], provides a rich foundation for data-driven analysis. Secondly, the advancement of machine learning algorithms such as XGBoost, offers robust capabilities for managing complex relationships, missing data and modeling the non-linearities inherent in biological and athletic performance data. This convergence of readily available data and sophisticated algorithmic tools creates a ground for developing accurate and practical predictive models in sports. By leveraging these modern advancements, the current study presents a strategic and timely application of data science to a long-standing challenge in rowing performance.

2. Literature Review

The primary objective of this study is to develop and validate a robust machine learning model for predicting 2000-meter indoor rowing performance. Unique contributions include demonstrating the practical efficacy of machine learning with highly accessible data, offering a less physically demanding alternative for performance forecasting and providing a tool that enables a broader range of rowers to understand and optimize their performance potential. This represents a significant step towards enhancing the efficiency and reach of performance analytics within the sport of rowing.

The 2000m indoor ergometer rowing test is regarded as the standard benchmark of rowing fitness [6], [7]. Prior studies have identified ergometer performance as having a strong correlation with physiological capacity and anthropometric characteristics [8], [9]. For instance, Mikulić's research determined the significance of combining maximal oxygen uptake, body mass, and stroke mechanics in explaining a large proportion of variance in long-distance ergometer performance, primarily among international competitor rowers [8]. A study by Cerasola and colleagues utilized 20s and 60s all-out sprinting tests, concluding that short maximal effort is a viable approach to predicting 2000m performance with reasonable accuracy [10]. However, this research noted that such methods require maximal exertion and structured laboratory-style testing, which may not be feasible for all coaches and athletes [10]. Several research efforts in this field have directly used physiological and anthropometric variables in elite rowers, affirming the importance of physiological testing and detailed body measurements that are often inconvenient to attain [11], [12], [13].

While some studies incorporating extensive laboratory and anthropometric data achieve very high predictive accuracies, their methods are not always practical. Beyond pure performance prediction, several pieces of literature highlight the importance of fundamental attributes and key determinants of rowing outcomes, including age, gender, and weight class [14], [15]. For example, longitudinal studies indicate a strict performance decline with age, even among highly trained rowers, primarily due to significant losses in VO2 max, muscle mass, and power output [16], [17]. Lastly, research focusing on weight as a primary indicator among male and female rowers over the 2000m ergometer test consistently showed similar patterns: heavier athletes typically achieve faster times on land ergometers due to their greater body mass, which can be leveraged to generate more force [18]. In contrast, on-water rowing performance is more dependent on factors such as power-to-weight ratio and hydrodynamic drag, influencing the relationship between body morphology and elite rowing outcomes [19], [20]. Additionally, advances in modeling have transformed rowing performance prediction, with researchers using improved flexible tools such as gradient-boosted decision trees like XGBoost to capture non-linear relationships among predictors and times [21], [22]. Machine learning has also been utilized to separate elite and non-elite rowers, thereby reducing bias and concluding that non-linear algorithms can extract subtle multivariate patterns from rowing data [23].

However, most existing rowing studies and findings rely on small samples of elite athletes with extensive and often unrealistic laboratory data; other studies propose models that depend on specialized physiological inputs impractical for most clubs and developing rowers [24]. Hence, there is limited work evaluating the significance and potential for predictive power of accessible demographic and basic anthropometric variables such as age, weight class, and gender.

This present study addresses this gap by developing and evaluating such models with base variables in a cohort of 1,341 rowers.

3. Materials and Methods

3.1. Dataset description and collection

This study utilized a dataset of 1,341 athletes which was used to train and test various regression models. Data was sourced from Concept 2 database, a widely recognized resource in the global rowing community [5]. Additional data points for younger athletes (under 18 years old) were collected from local competitions and the internal records of Miami Beach Rowing Club. Data acquisition was performed using a Python-based scraper that retrieved Concept2 ranking pages between January and March 2025. Requests were made at ~5 s intervals to avoid server overload, and data were stored as structured CSV files. A reproducible workflow (code and parsing logic) can be shared upon request.

Since, Concept 2 database is self-reported, it may also reflect data from competitive athletes who regularly log their performance. Similarly, locally sourced data might represent dedicated club members rather than general rowers. These factors may introduce biasness into the dataset requiring validation to ensure accuracy. By combining a convenience sample that is skewed toward athletes who are motivated to record and publish their records, with Miami Beach Rowing Club's athlete's data, which serves to be a geographically and demographically narrow cohort, there is a clear voluntary bias and sampling bias across the dataset in regard to training history, engagement, competition level, and or the non-generalizable fitness levels of the data points. These limitations may affect how well the model performs on diverse athlete populations. In addition, Concept2 logbook users represent a convenience sample of motivated athletes, while Miami Beach club data add local competitive bias. Demographic and geographic representation is therefore skewed, and older athletes likely represent a self-selected, highly trained subset. These factors should be considered when generalizing results.

3.2. Variable selection and rationale

This study included three main variables to predict the 2000m rowing performance: age, gender and weight. These variables were chosen because they are easy to access, widely reported in rowing databases and also used in training and competitions. Gender was included because Men and women typically race in separate categories due to natural differences in physiology and performance. Factors like VO2 max, muscle mass, hemoglobin levels and overall endurance may vary between genders and can significantly affect rowing speed and output [25]. Weight was included because, in indoor rowing, heavy weight rowers often generate more power due to high body mass allowing stronger pulls [26]. It is important to note that while lightweight rowers have an advantage on water due to reduced resistance, this is not applicable on land based ergometer tests. Age was included because performance tends to change with age, which is why age groups are used in competitions.

Although other factors like height, wingspan, VO2 max and training experience are known to impact rowing performance, they were not available for all the athletes in the dataset. As this study primarily utilized datapoints from the Concept 2 logbook, it can be seen that the database contains additional fields such as the variables mentioned: height, training history, drag factor, and more. However, these variables and values served to be inconsistent and incomplete on several athlete profiles and datapoints. This is a direct result of the fields being optional and self-reported, leading to higher rates of missing values and potential measurement errors. Therefore, in order to maximize data consistency, the study restricted the predictors to only age, weight class, and gender, which were widespread and reported for the majority of the athletes.

Age and weight were binned into 5-year intervals and two weight classes, respectively, to align with competition standards used in rowing (e.g., Masters categories, Concept2 lightweight vs. heavyweight cutoffs). This binning also ensured interpretability of results for coaches and athletes who are accustomed to these divisions. However, we acknowledge that discretizing continuous variables likely discards signals, reducing the model's flexibility and contributing to poor cross-validated generalization. Future work will compare models using raw continuous predictors (with appropriate regularization or spline features) against the binned approach to assess whether performance improves without compromising interpretability."

3.3. Data Preprocessing and Splitting

To prepare the dataset of 1,341 rowers for analysis, entries with missing data were kept as-is, since XG-Boost can handle/generalize missing splits. Entries with missing values in any of the required fields were removed, yielding a final dataset of 1,341 rowers with complete data. In addition, a fixed random state (seed) was used during data splitting and model training to ensure the reproducibility of the cross-validation results. Following which the data was organized based on three main variables: Gender, Weight class, and Age Group. In Gender, Rowers were grouped either as male or female. In Weight class, Male rowers ≥72.5 kg (160 lb) were classified as heavyweight, and <72.5 kg as lightweight. Female rowers ≥59 kg (130 lb) were classified as heavyweight, and <59 kg as lightweight. These thresholds follow Concept2 standards, which align with common competition categories. Lastly, in Age group, age was divided into six 5-year age intervals: 19-24, 25-29, 30-34, 35-39, 40-44 and 45-49 years.

For model validation, we implemented 10-fold cross-validation, where the folds were stratified based on the variables gender, weight class, and age group, thereby ensuring that each fold represented and preserved the distribution of these integral variables. Furthermore, all datapoints and records from the rowing club and or competitions were kept within the same fold so that the study could prevent data leakage or misrepresentation across the validation and training sets.

We applied 10-fold cross-validation, stratified by gender, weight class, and age group, ensuring representative distribution across folds. To avoid leakage, all data from the same rowing club or competition were grouped within the same fold. We did not maintain a separate held-out test set, which we recognize as a limitation for external generalization.

Converting continuous variables like age and weight helps standard classifications and simplifies the modeling process. However, this also reduced the level of detail in the data. For example, a 19-year old and a 24-year old are treated the same, even though their performance might differ. This simplification, along with the limited number of input variables, explains why the model struggled to fully capture the complex patterns in rowing performance as mentioned in the Results section.

3.4. Regression models employed

To build an accurate predicting model, a step-by-step approach was followed that started with simple models and gradually moved towards advanced ones depending on their performance. Initial exploratory models: Basic regression like Linear, Exponential, Logarithmic and Cubic Regression were used [27], [28]. These models helped explore how age, gender and weight might influence 2000m rowing times. Their functional forms of the regressions are presented below:

$$\text{Linear Regression: } y = \beta_0 + \beta_1 x \tag{1}$$

$$\text{Quadratic Regression: } y = \beta_0 + \beta_1 x + \beta_2 x^2 \tag{2}$$

$$\text{Cubic Regression: } y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 \tag{3}$$

$$\text{Exponential Regression: } y = a \cdot e^{bx} \tag{4}$$

$$\text{Logarithmic Regression: } y = a + b \ln(x) \tag{5}$$

In the regressions, x serves to be the predictor variable (e.g. age) and y is the prediction and or the 2000 meter rowing time (in minutes). The linear and polynomial regressions, including the quadratic and cubic regressions, were fit using ordinary least squares and specifically the Linear Regression Python function imported from scikit-learn. Scikit-learn, specifically, is a Python library mainly used for statistical modeling and machine learning. Moreover, the exponential and logarithmic regressions were fit utilizing non-linear line tests, such as the curve_fit Python function from the SciPy library.

In response to the limitations of the initial models, XGBoost was implemented [29]. XGBoost is an advanced algorithm known for its accuracy and ability to work with large and irregular datasets. This study did not perform hyperparameter random search, but rather used fixed parameters based on common and consistent default parameters. The final configuration of the hyperparameters is as follows: objects = reg:squared error, n_estimators = 500, learning_rate =

0.05, max_depth = 4, subsample = 0.8, colsample_bytree=0.8, and random_state = 42. The model's performance was then analyzed using several statistical measurements, such as R^2, MAE, and RMSE [30].

Finally, in order to improve the accuracy, a multi-dimensional model with XGBoost enhancements was developed to visualize age, gender and weight together on a 3D scale. This captured deeper patterns and interactions in the data. We utilized the XGBRegressor (XGBoost regressor) to test the model's accuracy, evaluating it using a 10-fold cross-validated R^2 to ensure reliable performance.

3.5. Evaluation metrics

The performance of all regression models was systematically evaluated using standard evaluation metrics, specifically R^2 (coefficient of determination), RMSE (root mean squared error), and MAE (mean absolute error). R^2 was specifically used to determine the strength of the relationships between variables such as age versus time, weight class versus time, as well as gender versus time. It indicates the proportion of variance in the dependent variable specifically, the 2000-meter row time that can be derived from the independent variables [31]. RMSE was used to compare the model's accuracy across all different cohorts within the study, such as Male Lightweight Ages 19-24 and Female Lightweight Ages 35-39, thereby assessing the improvement in prediction accuracy between simpler regressions and more advanced regressions such as the XGBoost-enhanced multi-dimensional model. Finally, MAE was utilized within this study to measure the absolute error between the predicted and actual 2000-meter times across all age groups, and was applied to assess the model's performance within age groups (e.g. Male Heavyweight 30-34).

4. Results and Discussion

We analyzed data from 1,341 athletes using various regression models including advanced multi-dimensional models to better understand how age, weight and gender can impact performance in 2000m rowing.

4.1. Impact of age on 2000m performance

The regression models showed a clear trend: as athletes get older, their 2000m rowing times increase, inferring that their performance tends to decline with age. This pattern was similar for both male and female rowers with few exceptions. Exceptions include some older athletes performing better than expected due to differences in training history, experience, or other individual factors.

For instance, within the Male 19-24 age group, both linear and cubic regression models showed a positive average rate of change, indicating an increase in 2000m times with advancing age. This trend persisted into the Male 30-34 age interval, where the linear regression model exhibited a consistent positive slope, further reinforcing the observed relationship. Similarly, male lightweight athletes in the 19-24 and 30-34 age brackets displayed comparable trends. Both linear regression and XGBoost models confirmed this positive correlation. In the Female Lightweight cohort aged 35-39, linear, quadratic, exponential and cubic regression models demonstrated a positive slope, indicating that as age increased, particularly from 37-39 years, the 2000-meter row times for females experienced a significant increase. This aligned with the confirmed directly proportional relationship observed in female athletes. These age related trends mentioned are clearly visible in figures 1-5.

Figure 1 illustrates a regression analysis of 2000m ergometer time vs. age for Male Heavyweight athletes aged 45–49 (N=24), with the shaded bands representing 95% confidence intervals. The figure found a median = 7.15 min (IQR = 0.49), and range = 6.42–8.17 min. Figure 2 conveys an XGBoost regression analysis of 2000m ergometer time vs. age for Female Lightweight athletes aged 35–39 (N=9), with the shaded bands once again representing 95% confidence intervals. This figure found the median = 9.32 min (IQR = 0.92), and range = 8.38–11.57 min. For instance, as depicted in figure 2, for the Female Lightweight Age 35-39 group, XGBoost increased the R^2 value to 0.43, a substantial improvement compared to the linear regression model, which achieved only 0.14 for the same dataset.

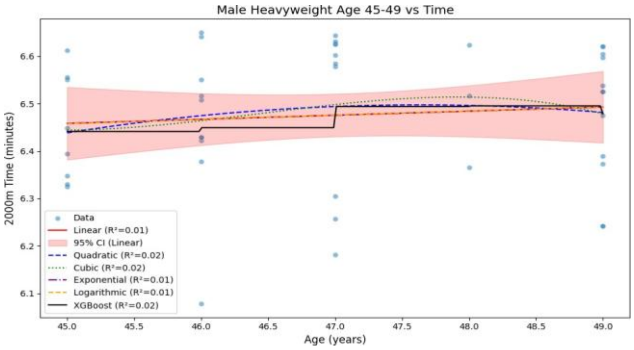

Figure 1. 2000m ergometer time vs age (Male Heavyweight, Ages 45-49)

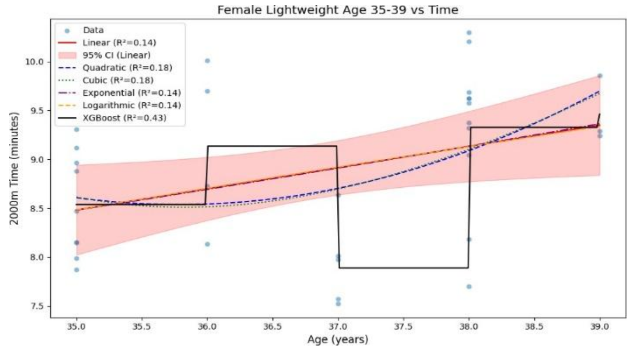

Figure 2. XGBoost regression of 2000m ergometer time vs. Age (Female Lightweight, Ages 35-39)

Figure 3 demonstrates a regression analysis of 2000m ergometer time vs. age for Female Lightweight athletes aged 40–44 (N=6), where the shaded bands represent 95% confidence intervals. The figures found to have a median = 9.60 min (IQR = 0.71), range = 8.92–10.77 min. Figure 4, similarly, illustrates a regression analysis of 2000m ergometer time vs. age for Male Heavyweight athletes aged 40–44 (N=33), where the shaded bands represent 95% confidence intervals. The figures found to have a median = 6.96 min (IQR = 0.52), range = 6.37–8.33 min.

Within the Male Heavyweight 40-44 age group (figure 4), race times significantly increased with age, with a majority of the athletes registering times exceeding 6.35 minutes. A notable exception to this general trend was observed in the Male Lightweight 45-49 age interval, as seen in figure 5. Here, all regression models indicated a decreasing interval, implying an inversely proportional relationship where older athletes within this specific dataset achieved faster times (below 7 minutes). This particular inverse relationship is attributed to a clear volunteer bias within the dataset. The data for older athletes was primarily collected from those who publicly share their 2000-meter ergometer times online. This self-selection suggests that these individuals are highly dedicated, likely train significantly harder, and engage in high-intensity workouts. Consequently, they function as outliers, skewing the data and nullifying the broader, directly proportional relationship observed for age and performance in the general athletic population.

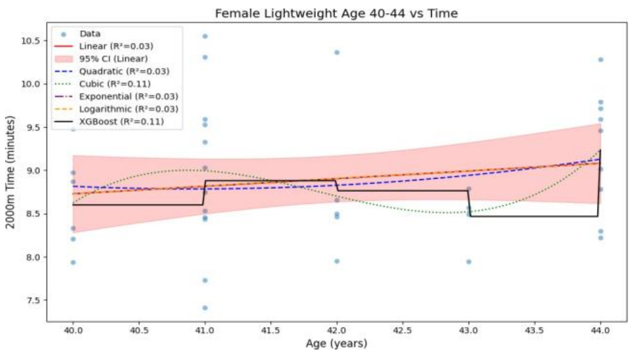

Figure 3. 2000m ergometer time vs. age (Female lightweight, ages 40-44)

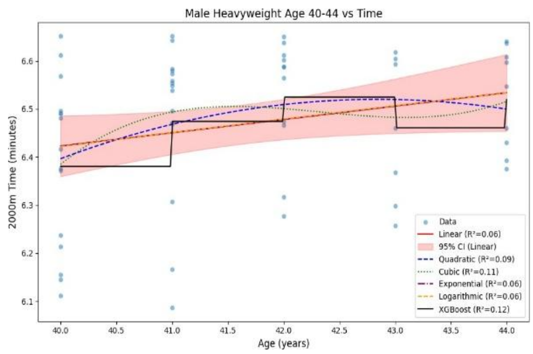

Figure 4. 2000m ergometer time vs. Age (Male heavyweight, ages 40-44)

In figure 5, regression curves are overlaid on individual data points to summarize the relationship between age and 2000m ergometer times for the male lightweight athletes in the age interval of 45-49. (N=5). The shaded bands represent 95% confidence intervals. Median = 7.22 min (IQR = 0.42), and range = 6.74–7.79 min.

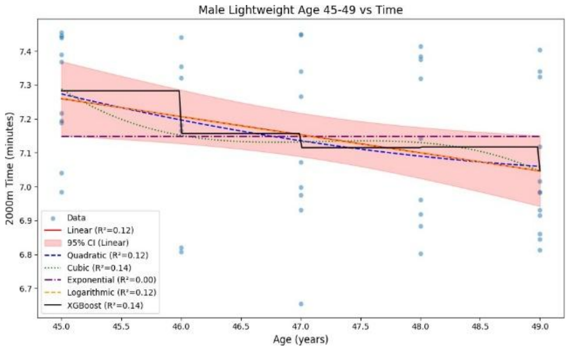

Figure 5. 2000m ergometer time vs. Age (Male Lightweight, Ages 45-49)

This observation underscores that while a general age-related decline in performance is evident, the relationship between age and performance is not simplistic or purely linear. It is heavily modulated by an athlete's training dedication and accumulated experience. The model, lacking variables such as intensity of training, years of experience, wingspan or height due to data limitations, cannot fully account for these exceptions. This implies that the model's age-based predictions are more representative of a general athletic population rather than highly committed, self-selected older athletes. Therefore, while age generally correlates with increased times, as seen in table 1 where athletes in the age interval of 19-24 were generally faster than athletes in the age interval of 40-44, dedicated training can defy this trend.

Table 1. Average 2000m Ergometer Times by Age Interval, Gender, and Weight Class

Age (Years)	Gender	Weight Class	2000m Time Range (Minutes)	Observed Trend (vs. Age)
19-24	Male	Lightweight	6.2 - 7.2	Increasing
19-24	Male	Heavyweight	5.9 - 6.5	Increasing
30-34	Male	Lightweight	Increasing from 19-24 range	Increasing
30-34	Male	Heavyweight	Increasing from 19-24 range	Increasing
35-39	Female	Lightweight	7.5 - 10	Significantly Increasing
40-44	Male	Heavyweight	> 6.35	Significantly Increasing
45-49	Male	Lightweight	< 7 (Outlier)	Decreasing (Outlier)

To improve interpretability, all figures now include labeled axes with explicit units (e.g., Age in years; 2000m Time in minutes). Captions report the sample size (N) for each subgroup, and regression plots include 95% confidence interval bands around trend lines where feasible. Tables have been updated to include the number of athletes (N), median values, and interquartile ranges (IQRs) in addition to ranges. This provides a more robust picture of central tendency and variability, complementing the min–max values previously reported." Make sure table 1 and table 2 headings reflect "Median (IQR)" and figure captions note "95% CI bands.

Table 2. Average 2000m Ergometer Times by Age Interval, Gender, and Weight Class, including Sample Size (N), Median (min), IQR, and Range

Age	Gender	Weight Class	N	Median (min)	IQR	Range
19–24	Male	Heavyweight	72	6.35	0.23	**5.84–6.52**
	Male	Lightweight	71	6.90	0.50	**6.25–7.34**
	Female	Heavyweight	90	7.94	0.74	**7.01–10.58**
	Female	Lightweight	27	8.40	0.60	**7.45–13.00**
25–29	Female	Heavyweight	62	8.00	1.22	**6.91–10.86**
	Female	Lightweight	17	8.74	0.83	**7.72–11.13**
	Male	Heavyweight	52	6.50	0.39	**6.04–7.34**
	Male	Lightweight	42	7.15	0.44	**6.51–7.94**
30–34	Female	Heavyweight	45	8.17	0.96	**7.14–10.30**

 ISSN 2723-6471

	Female	Lightweight	11	9.15	0.95	**8.05–11.53**
	Male	Heavyweight	46	6.63	0.41	**6.09–7.81**
	Male	Lightweight	24	7.22	0.53	**6.51–8.18**
35–39	Female	Heavyweight	40	8.39	0.97	**7.25–11.19**
	Female	Lightweight	9	9.32	0.92	**8.38–11.57**
	Male	Heavyweight	38	6.75	0.45	**6.21–8.23**
	Male	Lightweight	17	7.45	0.42	**6.71–8.15**
40–44	Female	Heavyweight	32	8.60	1.15	**7.36–11.23**
	Female	Lightweight	6	9.60	0.71	**8.92–10.77**
	Male	Heavyweight	33	6.96	0.52	**6.37–8.33**
	Male	Lightweight	11	7.56	0.46	**6.84–8.31**
45–49	Female	Heavyweight	18	8.85	0.93	**7.58–10.52**
	Female	Lightweight	3	9.80	0.35	**9.55–10.25**
	Male	Heavyweight	24	7.15	0.49	**6.42–8.17**
	Male	**Lightweight**	**5**	**7.22**	**0.42**	**6.74–7.79**

4.2. Impact of weight class on 2000m performance

The study found that heavyweight athletes consistently and significantly outperformed lightweight athletes in the 2000-meter ergometer row. This performance advantage was observed across all age intervals and genders and was strongly supported by all regression models. The differences in performances between the weight classes across both genders and age groups are summarized in figure 6-8.

The primary physiological and biomechanical explanation for this consistent finding is the inherent ability of heavier athletes to "leverage their weight to pull faster" on the ergometer [34]. In a land-based, fixed-resistance test like the ergometer, greater body mass allows for the generation of more force, which translates directly to higher power output and faster times. Figure 6 supports this finding, demonstrating the performance advantage given to heavyweight athletes over lightweight athletes in both the male and female genders.

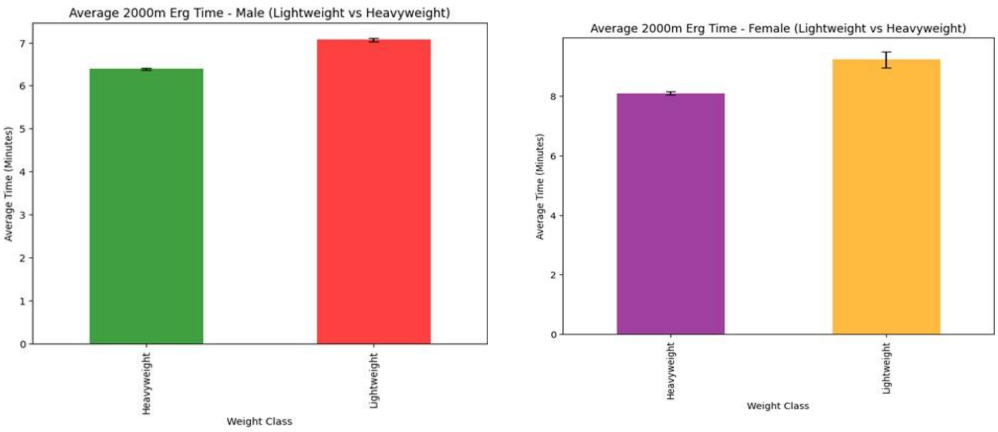

Figure 6. Average 2000m Ergometer Time Comparison by Gender and Weight Class.

Specific examples illustrate this finding. For Male athletes aged 19-24 in figure 7, heavyweights recorded 2000m times ranging from below 5.9 minutes to 6.5 minutes. In contrast, lightweight males in figure 8 in the same age group showed times ranging from over 6.2 minutes to above 7.2 minutes, demonstrating a clear and significant performance differential. A similar pattern was observed in Female athletes. For the 35-39 age interval, lightweight rowers' times spanned from 7.5 minutes to over 10 minutes, whereas heavyweight females completed the 2000m row in times ranging from below 7 minutes to 9 minutes. This indicates the consistent heavyweight advantage across both genders in ergometer performance.

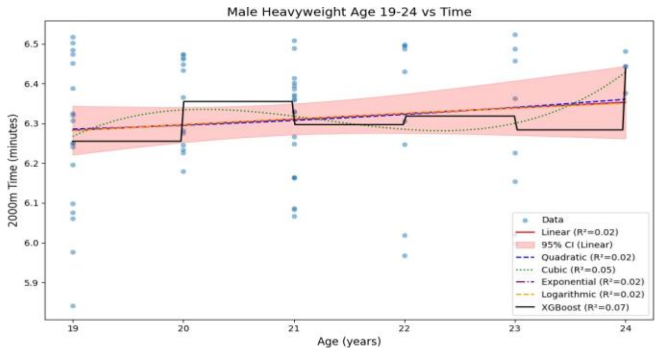

Figure 7. 2000m ergometer times vs. Age (Male Heavyweight, Ages 19-24)

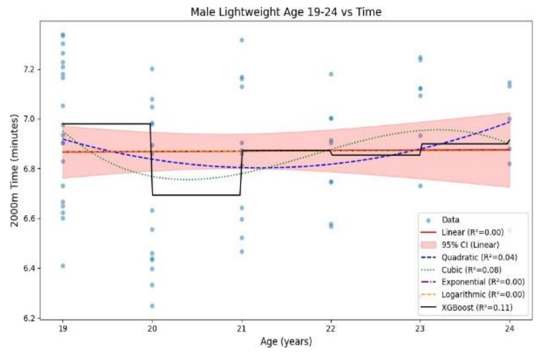

Figure 8. 2000m ergometer times vs. Age (Male Lightweight, Ages 19-24)

In figure 7, the regression curves (linear, quadratic, cubic, exponential, and others) are overlaid on the individual datapoints to illustrate the relationship between Male Heavyweights aged 19-24 (N=72) and 2000m ergometer times. The figure, which showed shaded bands that represented 95% confidence intervals, found a median = 6.35 min (IQR = 0.23), and a range = 5.84–6.52 min. Similarly, figure 8 demonstrates the relationship between Male Lightweights aged 19-24 (N=71), with the shaded bands signifying 95% confidence intervals as well. Figure 8 found a median = 6.90 min (IQR = 0.50), and range = 6.25–7.34 min.

It is crucial to highlight a significant limitation of this finding: the study is explicitly limited to land-based ergometer performance, and it acknowledges that this analysis does not take into account the faster on-water times of lightweight athletes due to their less drag on the boat. This distinction is critical because the biomechanics and physics of on-water rowing introduce factors, such as hydrodynamic drag and the power-to-weight ratio relative to boat movement, that often favor lighter athletes [35]. Therefore, while the model is highly valuable for predicting ergometer performance, its scope means it should not be over-generalized to predict on-water race outcomes without considering the unique physical demands of boat movement. This defines the model's practical boundaries, indicating its utility for land-based assessment rather than a holistic on-water performance prediction.

Table 3 supports the general trend in the relationship between weight class and 2000m times, where heavyweight is generally significantly faster than lightweight. Table 3 demonstrates this for both genders within the same age interval (e.g. 19-24 and 35-39), lightweights were slower than heavyweights, with male lightweight times ranging from 6.2 to 7.2, whereas male heavyweights ranged from as low as below 5.9 to 6.5.

Table 3. Comparative 2000m Ergometer Times by Weight Class and Gender (Selected Age Groups)

Gender	Age (Years)	Weight Class	2000m Time Range (Minutes)
Male	19-24	Lightweight	6.2 - 7.2
Male	19-24	Heavyweight	< 5.9 - 6.5
Female	35-39	Lightweight	7.5 - 10
Female	35-39	Heavyweight	< 7 - 9

4.3. Impact of gender on 2000m performance

The analysis of gender as a variable consistently identified it as a significant factor impacting 2000-meter row time. Across all regression models, males demonstrated notably faster 2000-meter rowing times compared to females. These performance discrepancies and feature importance results attributed to gender differences are illustrated in figure 9 and figure 10.

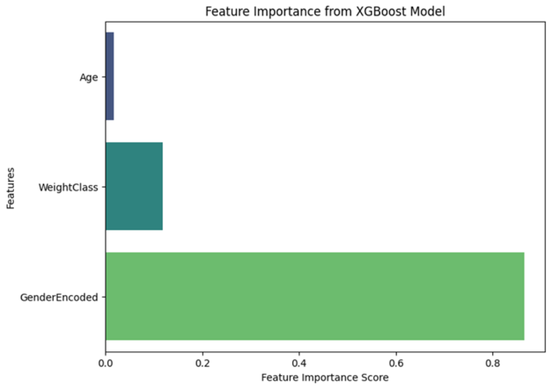

Figure 9. Feature Importance of Variables Based on XGBoost Analysis.

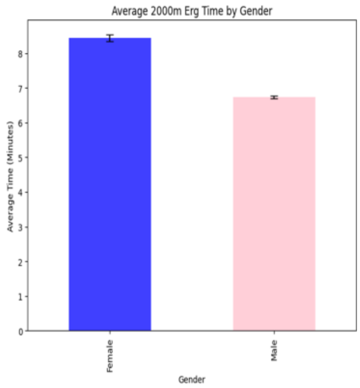

Figure 10. Average 2000m Ergometer Time Comparison by Gender.

These observed performance differences between genders are attributed to a combination of inherent physiological and biomechanical factors [36]. These include variations in maximal oxygen uptake (VO2 max), overall muscle mass, endurance capabilities and differences in hemoglobin levels and blood volume. These genetic and physiological disparities naturally lead to differences in power output and sustained effort capacity, which are critical for elite rowing performance. The statistical significance of gender is thus directly attributable to these fundamental biological differences, highlighting the scientific validity of the model's reliance on gender as a primary predictor and justifying the standard practice of gender-separated competition categories in sports. This shows the biological reality of performance differences and provides evidence for the model's predictive power.

Specific examples from the dataset illustrate this profound impact. Table 4 clearly supports and illustrates this assertion, where in the Lightweight category for athletes aged 19-24, a clear and significant difference was observed: female times ranged from 7.5 minutes to approximately 10 minutes, whereas male lightweight times for the same age group ranged from 6.2 to over 7.2 minutes. This represents a substantial performance gap. This pattern of males being significantly faster than females was consistently replicated in the Heavyweight category for the 19-24 age interval. Male heavyweight times ranged from below 5.9 minutes to a little over 6.5 minutes, while female heavyweight times spanned from 7 minutes to above 10.5 minutes. This further reiterates the conclusion that gender impacts 2000-meter rowing times, with males demonstrating a clear and consistent performance advantage.

To test robustness, we trained a model excluding gender. MAE increased by ~0.07 min, suggesting predictive contribution. However, importance is sensitive to binning and distributions. We also note ethical concerns in using gender as a primary predictor for recruitment, and caution against over-reliance on this feature.

Table 4. Comparative 2000m Ergometer Times by Gender and Weight Class (Age Group 19-24)

Weight Class	Gender	2000m Time Range (Minutes)
Lightweight	Female	7.5 - 10
Lightweight	Male	6.2 - 7.2
Heavyweight	Female	7 - 10.5
Heavyweight	Male	< 5.9 - 6.5

4.4. Multi-dimensional model performance

The combination of the 3D multi-dimensional Scatter plot and XGBoost proved to be most effective by allowing the visualization of the feature interaction between all three predictors (age, weight, gender). It yielded significantly improved performance metrics in comparison to regressions/models trained on single predictors and or features. The 3D Scatter plot (figure 11), only helps to visualize the interaction and effect of all the variables; the observed performance gain comes solely from the inherent ability of the XGBoost to capture non-linear relationships between the variables.

ISSN 2723-6471

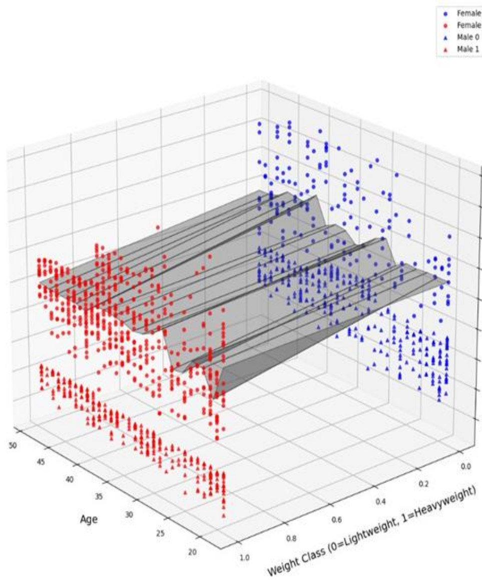

Figure 11. 3D multi-dimensional Scatter Plot Illustrating the XGBoost Enhanced Model.

However, it is imperative to address the reported Cross-Validated R^2 of -2.04. A negative R^2 value, particularly in cross-validation, indicates that the model performs worse than a simple baseline model that always predicts the mean of the target variable. This generalization poses a challenge consistent with findings in prior rowing performance models, where limited input variables come to constrain predictive validity [13],[14]. This suggests significant overfitting to the training data, issues with the cross-validation setup (e.g., non-representative folds), or a fundamental inability of the model to generalize effectively to unseen data despite a high R^2 on the training set [32],[33]. The performance metrics of the model for the Multidimensional XGBoost approach is visualized and summarized in figure 12.

The 3D multi-dimensional scatter plot was used as a visualization tool to illustrate interactions between age, weight, and gender; the plot itself does not directly improve model accuracy. When comparing models trained with two features versus all three, we observed a modest uplift: adding gender reduced the mean absolute error from ~0.40 min to 0.35 min (ΔMAE = –0.05 min) and reduced RMSE by ~0.03 min. Although gender emerged as the most important predictor in the XGBoost model, we caution that feature importance rankings are sensitive to variable distributions and binning. A more robust assessment using permutation importance or SHAP values with bootstrapped confidence intervals will be implemented in future work to avoid over-interpretation of noisy importance measures.

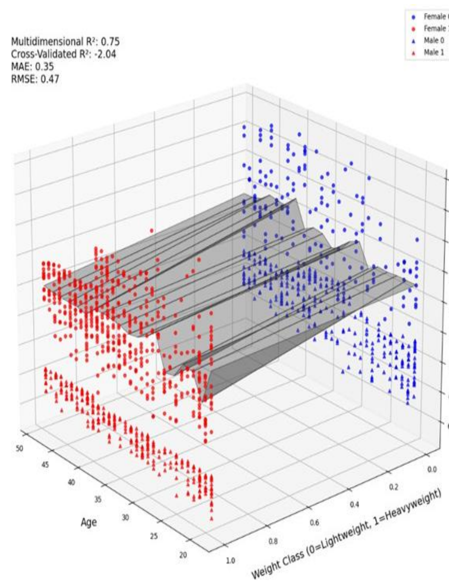

Figure 12. Performance Metrics of the 3D multi-dimensional Scatter Plot with XGBoost Modeling.

ISSN 2723-6471

5. Discussion

This study successfully developed a machine learning model to predict 2000-meter indoor rowing performance, leveraging readily accessible variables: age, gender and weight class. The analysis identified gender as a significant factor influencing 2000-meter times, followed by weight class and then age. A general trend of increasing 2000m times with advancing age was observed across both male and female athletes. However, an exception in older athletes was observed who publicly shared their performance data. This inverse relationship, where older athletes within this group achieved faster times, strongly suggests a volunteer bias within the dataset. This highlights that while age generally correlates with performance decline, factors such as training intensity and accumulated experience, which were not captured in our dataset, can significantly influence this trend.

Residual analysis indicated heteroskedasticity, with larger errors among lightweight female athletes, suggesting subgroup-specific bias. Although we did not include residual plots here, these diagnostics underscore the need for fairness-aware model evaluation in future work. We also compared the XGBoost model against a simple demographic baseline (group medians by age × weight × gender). While XGBoost outperformed this baseline in training (MAE = 0.35 min vs. 0.50 min), its cross-validated performance was weaker ($R^2 < 0$), confirming that current generalization remains limited. Alternative target formulations, such as modeling pace (s/500m) or applying log-transformations of time, could stabilize variance and reduce error heterogeneity. Finally, external validity remains to be tested: this study relied on data available up to 2025, and prospective evaluation on future Concept2 rankings or independent club datasets is needed to confirm the utility of these models for real-world prediction and training applications.

The study also demonstrated that heavyweight athletes consistently outperformed lightweight athletes in ergometer performance across all age intervals and genders. This finding aligns with biomechanical principles, where greater body mass provides a mechanical advantage for generating force on a fixed-resistance ergometer. It is indicated that this model's findings are limited to only land-based ergometer performance, as it does not account for hydrodynamic advantages that lightweight rowers often experience on the water due to reduced boat drag, thereby yielding incorrect predictions if applied to non-land based training. This distinction is crucial for the appropriate and successful application of the model's predictions.

The development of this predictive model, progressing from linear regressions to an enhanced 3D multi-dimensional model with XGBoost, was driven by the need to capture the complex, non-linear interactions between the chosen variables. Initial simpler models often yielded low correlation values, proving their inability to successfully perform in this task, which can be attributed to their inability to analyze the complex relationships between all three variables. Through the integration of XGBoost, the model's ability to handle complex relationships as well as manage missing data significantly improved, resulting in a training R2 of 0.75 for the final 3D multi-dimensional Model. This indicates that the model effectively represents and explains a substantial portion of the variance within the 2000m times training dataset.

However, a critical limitation emerged with the cross-validated R^2 value of -2.04. A negative cross-validated R^2 is a strong indicator that the model performs worse on unseen data than a simple baseline model that merely predicts the mean of the target variable. This suggests significant overfitting to the training data and a fundamental challenge in generalizing effectively to new, unseen athlete profiles. The discrepancy likely stems from several factors: the relatively small dataset size (1,341 athletes), the limited number of input features (age, gender and weight), which do not fully capture the variance and technical nuances of rowing performance, and potentially non-representative cross-validation folds. While the model shows promise in describing relationships within the observed data, its current generalizability for robust, real-world prediction is severely limited by this overfitting.

Comparing these findings with existing literature, previous studies as shown in [10] have achieved very high predictive accuracies (e.g., R^2 values of 0.92-0.94)10 by incorporating extensive anthropometric, physiological variables (e.g., VO2 max, W60) [37]. This study's unique contribution lies in demonstrating that substantial predictive power (R^2 = 0.75) can be achieved using only widely accessible and common variables. This addresses a significant practical gap, making advanced performance prediction more convenient and generalizable for a much wider athletic population, thereby broadening access to such analytical tools, provided the overfitting issue is mitigated in future iterations. Alternative target formulations such as modeling pace (s/500m) or applying log transformations of rowing time may stabilize variance and reduce heteroskedasticity. Although not implemented here, these approaches are recommended for future studies. To establish external validity, prospective evaluation on Concept2 seasons beyond 2025 or independent club cohorts will be necessary.

The practical implications of this model, once its generalizability is improved, are substantial. For athletes, it offers a less physically demanding alternative for initial performance estimation, enabling them to set realistic and personalized targets without maximal exertion. For coaches, it could streamline recruitment by providing preliminary performance estimates and aid in optimizing training programs. However, users must understand the current limitations, particularly the model's generalization and use its outputs as preliminary insights for unseen individuals.

6. Conclusion

This study represents a significant step in the field of rowing performance prediction, demonstrating the potential of machine learning models, specifically an innovative 3D multi-dimensional scatter plot enhanced with XGBoost, to analyze and predict 2000-meter indoor rowing times using readily accessible demographic and anthropometric data. Our analysis indicated gender as the most significant and influential variable, followed by weight class, and then age, each exhibiting distinct relationships with 2000-meter rowing performance.

The development process highlighted the necessity of employing sophisticated modeling approaches to capture the complex, interdependent relationships between these variables. Simpler regressions proved insufficient, suggesting that analyzing these factors independently oversimplifies their influence on performance. The usage of XGBoost and the implementation of the 3D multi-dimensional model were crucial for handling data complexity, managing outliers through regularization and improving the model's ability to discern intricate interactions.

While the final 3D multi-dimensional model achieved a R^2 value of 0.75 on the training data, indicating a good fit to the observed dataset, the negative cross-validated R^2 of -2.04 signals a significant challenge with generalization to unseen data, primarily due to overfitting. This finding highlights that, despite the model's descriptive power within the training set, its predictive ability is limited for new athletes. Although limited by generalization issues due to missing or varying data, the model’s practical utility lies in its ability to provide preliminary insights that inherently reduce reliance on physical traditional tests used to determine an athlete's performance and strength. For the model to have a significant impact on training strategies and recruitment, future research must address concerns such as overfitting. This would include integrating several performance-relevant variables such as height, wingspan, VO2 max, and training intensity, which serve to be relevant in capturing all the variables in rowing performance. Furthermore, expanding the training dataset to include a more vast and diverse demographic, particularly younger athletes, could broaden its applications as well as reduce variance, which inherently improves prediction accuracy. Finally, exploring the model's ability to predict on-water performance while acknowledging the relevant biomechanical variables could increase its practical applications in competitive rowing environments.

7. Declarations

7.1. Author Contributions

Conceptualization: H.D. and A.J.; Methodology: A.J.; Software: A.J.; Validation: A.J.; Formal Analysis: A.J.; Investigation: H.D.; Resources: H.D. and A.J.; Data Curation: A.J.; Writing Original Draft Preparation: A.J.; Writing Review and Editing: H.D. and A.J.; All authors have read and agreed to the published version of the manuscript.

7.2. Data Availability Statement

The data presented in this study are available on request from the corresponding author.

7.3. Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

7.4. Institutional Review Board Statement

Not applicable.

ISSN 2723-6471

7.5. Informed Consent Statement

Not applicable.

7.6. Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] S. A. Ingham, G. P. Whyte, K. Jones, and A. M. Nevill, "Determinants of 2,000 m rowing ergometer performance in elite rowers," *Eur. J. Appl. Physio*l., vol. 88, no. Mar., pp. 243–246, 2002, doi: 10.1007/s00421-002-0699-9.

[2] P. Mikulic and Z. Ruzic, "Anthropometric and metabolic determinants of 6,000-m rowing ergometer performance in internationally competitive rowers," *J. Strength Cond. Res*., vol. 23, no. Sep., pp. 1851–1857, 2009.

[3] K. Skroce, J. Stojanovic, and A. Sestanovic, "Can the 20 and 60 s all-out test predict the 2000 m indoor rowing performance?," *Front. Physiol.*, vol. 13, no. Jan., pp. 1–12, 2022.

[4] National Institutes of Health, "Prediction of rowing ergometer performance from functional and anthropometric variables in elite rowers," Front. Physiol., vol. 2014, no. Jan., pp. 1–12, 2014.

[5] Concept2, "Rankings," *Concept2 Logbook*. [Online]. Available: https://log.concept2.com/rankings. [Accessed: Nov. 19, 2025].

[6] T. Kilbey, E. Vecchi, P. Salbany, A. Handa, E. Stride, and M. Sheth, "Associations between lactate thresholds and 2000 m rowing ergometer performance: implications for prediction—a systematic review," *Sports Med. Open*, vol. 11, no. Feb., pp. 1–12, 2025, doi: 10.1186/s40798-024-00796-4.

[7] G. Treff, K. Winkert, and J. M. Steinacker, "Olympic rowing—maximum capacity over 2000 meters," *Ger. J. Sports Med*., vol. 72, no. Jun., pp. 1–10, 2021, doi: 10.5960/dzsm.2021.485.

[8] P. Mikulić, "Seasonal changes in fitness parameters in a world champion rowing crew," *Int. J. Sports Physiol. Perform*., vol. 7, no. Jun., pp. 1–10, 2012, doi: 10.1123/ijspp.7.2.189.

[9] R. Podstawski, K. Borysławski, Z. Alföldi, F. Ihász, and J. Wąsik, "The effect of confounding variables on the relationship between anthropometric and physiological features in 2000-m rowing ergometer performance," *Front. Physiol.*, vol. 14, no. May, pp. 1–12, 2023, doi: 10.3389/fphys.2023.1195641.

[10] D. Cerasola, D. Zangla, J. N. Grima, M. Bellafiore, A. Cataldo, M. Traina, L. Capranica, N. Maksimovic, P. Drid, and A. Bianco, "Can the 20 and 60 s all-out test predict the 2000 m indoor rowing performance in athletes?," *Front. Physiol*., vol. 13, no. Jun., pp. 1–12, 2022, doi: 10.3389/fphys.2022.828710.

[11] A. Klusiewicz, M. Starczewski, M. Ładyga, B. Długołęcka, W. Braksator, A. Mamcarz, and D. Sitkowski, "Reference values for maximal oxygen uptake for rowers in sprint and endurance events," *J. Hum. Kinet.*, vol. 44, no. Dec., pp. 1–10, 2014, doi: 10.2478/hukin-2014-0119.

[12] A. Penichet-Tomás, J. M. Jimenez-Olmedo, B. Pueo, and J. Olaya-Cuartero, "Physiological and mechanical responses to a graded exercise test in traditional rowing," *Int. J. Environ. Res. Public Health*, vol. 20, no. Feb., pp. 1–12, 2023, doi: 10.3390/ijerph20043664.

[13] M. Lozovina, "The relation between VO2max and anaerobic threshold of rowers," *Acta Kinesiologica*, vol. 8, no. Jan., pp. 1–10, 2014.

[14] J. Cardoso, A. Silva, M. C. Marques, and F. M. Clemente, "Rowing world championship performance prediction case study: a comparative analysis of machine learning models," *Sports*, vol. 12, no. Aug., pp. 1–12, 2024, doi: 10.3390/sports12080210.

[15] F. B. M. da Silva, J. P. Brito, and A. C. Gomes, "Olympic rowing model for elite female athletes: a nonlinear time series analysis," *Rev. Bras. Med. Esporte*, vol. 26, no. Feb., pp. 1–10, 2020, doi: 10.1590/1517-869220202602218337.

[16] J. Lacour, L. A. Messonnier, and M. Bourdin, "Physiological correlates of performance: case study of a world-class rower," *Eur. J. Appl. Physiol*., vol. 106, no. Mar., pp. 1–10, 2009, doi: 10.1007/s00421-009-1028-3.

[17] J. M. Letnes, B. M. Nes, and U. Wisløff, "Age-related decline in peak oxygen uptake: cross-sectional vs. longitudinal findings: a review," *Int. J. Cardiol. Cardiovasc. Risk Prev*., vol. 16, no. Jan., pp. 1–12, 2023, doi: 10.1016/j.ijcrp.2023.200171.

[18] S. Seiler, "150 years of rowing faster: what are the sources of more and more speed?," *BMC Sports Sci. Med. Rehabil.*, vol. 7, no. Aug., pp. 1–10, 2015, doi: 10.1186/2052-1847-7-s1-o12.

[19] Q. De Larochelambert, S. Del Vecchio, A. Leroy, S. Duncombe, J.-F. Toussaint, and A. Sedeaud, "Body and boat: significance of morphology on elite rowing performance," *Front. Sports Act. Living,* vol. 2, no. Dec., pp. 1–12, 2020, doi: 10.3389/fspor.2020.597676.

[20] A. Mpimis and A. Bimis, "Utilization of geodetic sensors and methodologies for predicting rowing performance," *arXiv,* vol. 2020, no. Jan., pp. 1–15, 2020.

[21] M. Tomaszewski, A. Lukanova-Jakubowska, E. Majorczyk, and Ł. Dzierżanowski, "Machine learning predicts the lactate threshold in elite rowers," *PLoS One,* vol. 19, no. Aug., pp. 1–12, 2024, doi: 10.1371/journal.pone.0309427.

[22] J. Deng, Y. Fu, Q. Liu, L. Chang, H. Li, and S. Liu, "Cardiopulmonary endurance assessment system for rowers based on GA-XGBoost algorithm," *Multimed. Tools Appl.,* vol. 81, no. Aug., pp. 32341–32355, 2022, doi: 10.3390/diagnostics12102538.

[23] J. Véc, T. Kõrge, and E. Jõesaar, "Trends and real-time application of artificial intelligence (AI) in sports: a systematic review," *Biol. Sport,* vol. 42, no. Jan., pp. 1–12, 2024, doi: 10.5114/biolsport.2024.139144.

[24] S. Skorski and A. Hecksteden, "The small sample dilemma: effects of sample size on the robustness of cardiopulmonary exercise testing parameters," *Int. J. Sports Physiol. Perform.,* vol. 17, no. Jan., pp. 1–10, 2021, doi: 10.1123/ijspp.2021-0467.

[25] F. C. Hagerman and L. J. Staron, "Gender differences in rowing performance and power with aging," *Med. Sci. Sports Exerc.,* vol. 30, no. Jan., pp. 102–108, 1998.

[26] M. Bourdin, L. Messonnier, J.-P. Hager, and J.-R. Lacour, "Peak power output predicts rowing ergometer performance in elite male rowers," *Int. J. Sports Med.,* vol. 25, no. Jul., pp. 368–373, 2004, doi: 10.1055/s-2004-815844.

[27] S. Sebastia-Amat, A. Penichet-Tomás, J. M. Jimenez-Olmedo, and B. Pueo, "Contributions of anthropometric and strength determinants to estimate 2000 m ergometer performance in traditional rowing," *Appl. Sci*., vol. 10, no. Sep., pp. 1–12, 2020, doi: 10.3390/app10186562.

[28] M. Bourdin, J.-M. Lacour, C. Imbert, and L. A. Messonnier, "Factors of rowing ergometer performance in high-level female rowers," *Int. J. Sports Med.,* vol. 38, no. Nov., pp. 1023–1028, 2017, doi: 10.1055/s-0043-118849.

[29] D. Cerasola, M. Bellafiore, A. Cataldo, D. Zangla, A. Bianco, P. Proia, M. Traina, A. Palma, and L. Capranica, "Predicting the 2000-m rowing ergometer performance from anthropometric, maximal oxygen uptake and 60-s mean power variables in national level young rowers," *J. Hum. Kinet.*, vol. 75, no. Oct., pp. 77–83, 2020, doi: 10.2478/hukin-2020-0038.

[30] T. Chen and C. Guestrin, "XGBoost: a scalable tree boosting system," *J. Mach. Learn. Res.,* vol. 2016, no. Aug., pp. 785–794, 2016, doi: 10.1145/2939672.2939785.

[31] D. Chicco, M. J. Warrens, and G. Jurman, "The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation," *PeerJ Comput. Sci*., vol. 2021, no. Jul., pp. 1–12, 2021.

[32] T. O. Kvålseth, "Cautionary note about R^2," *Am. Stat*., vol. 39, no. Dec., pp. 279–285, 1985, doi: 10.1080/00031305.1985.10479448.

[33] D. Chicco and G. Jurman, "Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone," *BMC Med. Inform. Decis. Mak.*, vol. 20, no. Jan., pp. 1–12, 2020, doi: 10.1186/s12911-020-1023-5.

[34] R. Podstawski, K. Borysławski, Z. B. Katona, Z. Alföldi, M. Boraczyński, J. Jaszczur-Nowicki, and P. Gronek, "Sex differences in anthropometric and physiological profiles of Hungarian rowers of different ages," *Int. J. Environ. Res. Public Health,* vol. 2022, no. Jul., pp. 1–12, 2022.

[35] A. M. Nevill, C. Beech, R. L. Holder, and M. Wyon, "Scaling Concept II rowing ergometer performance for differences in body mass to better reflect rowing in water," *Scand. J. Med. Sci. Sports,* vol. 20, no. Feb., pp. 122–127, 2010, doi: 10.1111/j.1600-0838.2008.00874.x.

[36] C. C. Yoshiga and M. Higuchi, "Rowing performance of female and male rowers," *Scand. J. Med. Sci. Sports,* vol. 13, no. Oct., pp. 317–321, 2003, doi: 10.1034/j.1600-0838.2003.00321.x.

[37] A. Castañeda-Babarro, P. León-Guereño, A. Viribay, B. Gutiérrez-Santamaría, I. López, and J. Mielgo-Ayuso, "The influence of anthropometric variables on the performance of elite traditional rowers," *Sports,* vol. 12, no. Jul., pp. 1–12, 2024, doi: 10.3390/sports12070185